Revista Catarinense da CIÊNCIA CONTÁBIL
ISSN 2237-7662. Florianópolis, SC, v. 24. 1-31. e3526. 2025
https://doi.org/10.16930/2237-7662202535262
Available at http://revista.crcsc.org.br

CREDIT RISK MANAGEMENT THROUGH THE USE OF MACHINE LEARNING: THE CASE OF BANCO BS2

ALEX GIOVANI DE ASSIS[1]
Faculdade Fipecafi
https://orcid.org/0009-0005-7070-362X
alexgiovanideassis@gmail.com

SONIA ROSA ARBUES DECOSTER
Faculdade Fipecafi
https://orcid.org/0000-0002-0081-347X
sonia.decoster@fipecafi.org

ABSTRACT

Credit risk has played a central role in several global financial crises over the past three decades. An increasingly complex and interconnected financial landscape makes risk management essential for the stability and growth of financial institutions. This case study aims to analyze the use of machine learning specifically, the Gradient Boosting Decision Tree (GBDT) algorithm in a predictive model that combines significant financial and non-financial variables and incorporates credit bureau inquiries into Banco BS2's credit risk management process. The goal is to achieve greater accuracy in decision-making and improvements in risk mitigation. The F1 metric, employed as a measure of the model's precision, shows a superior value of 0.77 when compared with the model used by Serasa. Since 2022. the continuous monitoring capability offered by this predictive model has provided BS2 Bank with a real-time view of the financial health of its customer base, thereby facilitating the implementation of more assertive policies. Furthermore, the default rate among Banco BS2's corporate clients, as recorded by BCB-CADOC (2024), has been on a decline following the implementation of the new GBDT-based model. This study contributes to promoting innovation and competitiveness within financial institutions by encouraging transparency and strengthening the confidence of investors, stakeholders, and regulators such as the Central Bank through the adoption of Artificial Intelligence (AI) tools that detect credit risks early and help prevent systemic crises.

Keywords: Credit Risk Management. Default. Predictive Model. Machine Learning. AI.

Edited in Portuguese and English. Original version in Portuguese.

[1] **Correspondence Address:** R. Maestro Cardim, 1170 | Bela Vista | 01323-001 | São Paulo/SP | Brazil.

Received on: 08/19/2024**. Revised on**: 12/20/2024. **Accepted on**: 01/21/2025 by Prof. Dr. Rogério João Lunkes (Editor-in-Chief). **Published on:** 02/18/2025.

1 INTRODUCTION

Credit risk, as highlighted by Lassance and Ternoski (2021), refers to the probability of default on a financial obligation, which may result in substantial losses for financial institutions. Over the past decades, studies have demonstrated the importance of credit risk management for global financial stability, considering the impacts of systemic risks that defaults may impose on the overall financial market (Reinhart et al., 2020).

According to Dewasari et al. (2024), the increasing complexity of financial systems and heightened economic volatility necessitate robust credit analysis methodologies to mitigate risks. Traditional risk assessment models cannot keep pace with the accelerated rate driven by the growing volume, velocity, and complexity of transactions and financial data (Rahmani et al., 2023). Consequently, credit management demands predictive and technological approaches that enable financial institutions to make informed and effective decisions (Khemakhem & Boujelbene, 2018).

In the Brazilian context, credit bureaus are particularly significant for credit analysis because they provide detailed information on consumers' financial history and payment behavior, especially in a market where access to reliable financial data can mean the difference between a secure credit granting and a risky exposure (Serasa Experian, 2023). These bureaus serve as trustworthy sources to assist in assessing clients' repayment capabilities and supporting informed decision-making (Oliveira & Santos, 2022). The Central Bank's CMN Resolution No. 5.037, dated September 29. 2022. establishes "the foundation for data-sharing agreements with five credit bureau database managers, aiming to contribute to the expansion of credit access for Brazilians at a more affordable cost" (Banco Central do Brasil, 2022).

According to Fosu et al. (2023), credit bureaus provide substantial support in the decision-making process regarding credit approvals by generating a credit score that evaluates the client; thus, institutions can approve, adjust the granted volume, or deny credit. In developing countries, private-sector management of bureaus often proves more efficient compared to public administration (Oliveira & Santos, 2022).

As noted by Louzada et al. (2016), the necessity for effective risk management has driven financial institutions to pursue continuous improvements in credit analysis techniques, leading to the development and application of numerous quantitative models. The utilization of credit scoring methods has grown substantially over the past decade due to increased data access, enhanced computational power, regulatory requirements, and the demand for economic growth (Demirgüç-Kunt et al., 2017). In this context, innovations in machine learning and artificial intelligence are presented as essential tools for constructing predictive models that increase accuracy, enhance assertiveness, and facilitate prevention (Lessman et al., 2015).

As described by Provost and Fawcett (2001), the impact of systemic risks on the global economy underscores the importance of supervised learning models in detecting patterns of financial behavior that may predict default. The Basel III Accord, introduced in 2013. demonstrated the Basel Committee's increased commitment to encouraging more sophisticated models for standard credit risk calculation (Bank for International Settlements, 2024). As the market becomes increasingly competitive, dynamic, and interoperable, technologies such as artificial intelligence and machine learning are essential for the practical application of emerging technologies in monitoring financial risks, highlighting the importance of adapting credit models in scenarios of uncertainty and rapid chance (Mashrur et al., 2020).

The application of these advanced models, such as decision trees and boosting techniques, facilitates the adaptation of financial institutions to global economic changes and increases resilience in the face of crises (Reinhart et al., 2020). Additionally, according to the literature review conducted by Montevechi et al. (2024), recent advancements in machine learning models enable the handling of large datasets with enhanced predictive power, such that algorithms

including decision trees, Support Vector Machines (SVM), and neural networks demonstrate potential in refining the credit risk modeling process. However, as Montevechi et al. (2024) also emphasize, when it comes to classifier models, there is sufficient evidence to assert that no single option is considered the best.

AI-driven risk management extends its reach through credit risk assessment, market risk prediction, fraud detection, and compliance management (Rahman et al., 2021). The use of machine learning, through algorithms such as GBDT (Gradient Boosting Decision Tree), underscores how technological innovation can enhance the resilience of financial institutions and support long-term global economic stability (Lessman et al., 2015). According to Zhou et al. (2019), decision tree-based algorithms like GBDT, XGBoost, and LightGBM are among the most advanced machine learning algorithms developed in recent years, as they have achieved the expected predictive outcomes in tasks involving imbalanced data.

Moreover, the accuracy of a credit model depends not only on the algorithm used but also on the selection of the appropriate variables. In this regard, GBDT stands out by enabling the identification of which variables most significantly impact default prediction, thereby enhancing the model's ability to correctly classify risk profiles (Chen & Guestrin, 2016). In light of the aforementioned context, the objective of this study is to analyze the use of machine learning specifically the Gradient Boosting Decision Tree (GBDT) algorithm in credit risk management at Banco BS2. which has been utilizing AI since 2022. The guiding research question for the study is: **How can the use of a predictive model based on a machine learning algorithm, specifically the Gradient Boosting Decision Tree (GBDT), assist in preventing credit risk?**

This study aims to contribute to research on default risk management by promoting a more responsible credit granting process and employing a machine learning algorithm to reveal the predictors of risk. It also seeks to facilitate the dissemination of knowledge regarding machine learning tools within the Brazilian financial market scenario specifically, the Gradient Boosting Decision Tree (GBDT) algorithm especially given the nascent volume of studies using this algorithm in credit risk analysis. The remainder of this article is composed of five additional sections. Section 2 provides a detailed literature review on global credit risk, credit granting in Brazil through credit bureaus, and the use of machine learning in credit risk analysis. The methodological procedures employed in the study are described in Section 3. Section 4 presents and interprets the results. Finally, Section 5 offers concluding remarks, discusses the study's limitations, and provides recommendations for future research, followed by the bibliographic references.

2 THEORETICAL FRAMEWORK

2.1 Credit Risk Worldwide

Credit risk, according to Reinhart et al. (2020), is one of the key factors in analyzing and preventing global financial crises, necessitating constant monitoring and adjustments by regulators and financial institutions. Understanding past events aids in developing predictive models that bolster economic and institutional resilience (Reinhart et al., 2020). In this regard, one can trace the roots back to the Asian financial crisis of 1997, which, as noted by Agarwal and Vandana (2022), underscored the vulnerability of emerging economies in the face of excessive borrowing and systemic weaknesses. Similarly, as Calvo (2008) explains, the Russian financial crisis of 1998 was characterized by a series of adverse factors; the Russian government's inability to stabilize the economy and finance its debt led to a default and the devaluation of the ruble, thereby reinforcing the importance of effective sovereign credit risk management (Dabrowski, 2023).

The dot-com crisis, which occurred in the early 2000s, focused on the stock market for technology companies and was characterized by the collapse of many overvalued internet startups

(Rizvi et al., 2015) and by the need for greater diligence in granting credit to emerging companies (Rizvi et al., 2015). The financial crisis of 2008. triggered by the issuance of subprime mortgages in the United States, led to one of the largest global recessions in recent history and resulted in the collapse of major financial institutions, thereby demonstrating systemic failures in credit risk management (Mian & Sufi, 2009). This crisis spurred the Basel Accords, which aimed to strengthen credit granting practices and risk management, and encouraged the development of more sophisticated models for the evaluation and mitigation of credit risk (*Bank for International Settlements*, 2024).

According to Freeman et al. (2017), risk management is crucial for financial stability, particularly in the banking industry, where the risk of default can jeopardize an institution's sustainability. Historical analyses of default events underscore the importance of regulatory practices and, above all, the development and adoption of advanced predictive technologies for risk management, as highlighted by Rahman et al. (2021). Similarly, Dewasiri et al. (2024) found that financial institutions employing artificial intelligence and machine learning have demonstrated a superior ability to adapt to risks and make proactive, preventive decisions in volatile environments.

2.2 Credit Granting in Brazil Through Consultations with Credit Bureaus

In Brazil, credit granting is regulated by the Central Bank of Brazil (BCB), which utilizes credit bureaus to consolidate customer data (BCB, 2023). This approach facilitates more accurate risk analyses and helps preserve the liquidity of the financial system (Mendonça & Deos, 2020). Additionally, CMN Resolution No. 5.037 of 2022 establishes the Credit Information System (SCR), which comprises data submitted to the Central Bank of Brazil regarding credit operations, and seeks to promote the exchange of information among financial institutions and other entities (BCB, 2022).

Each credit bureau employs a scoring methodology, known as a credit score, which takes into account customers' financial histories; lower scores indicate a higher risk of default, and this analysis is fundamental to ensuring the integrity of credit operations in the financial market (Oliveira & Santos, 2022). Credit granting is supported by consultations with credit bureaus, which provide essential data for assessing the credit risk of both individuals and companies (Mendonça & Deos, 2020). According to Sfeir (2023), the Brazilian financial market frequently cross-references internal analyses with information supplied by credit bureaus leveraging data such as the Positive Registry and notes that innovation in the credit bureau sector, including integration with Open Banking, has enabled financial institutions to expand their analytical capabilities, offer more personalized products, and enhance the consumer experience.

The use of these scores by credit bureaus enables a standardized decision-making approach, while also facilitating access to the essential data needed for setting interest rates and credit limits (Oliveira & Santos, 2022). The adoption of these practices, as regulated by the Central Bank of Brazil (BCB, 2023), is intended to promote the safety and stability of the Brazilian financial system (BCB, 2023). This classification system allows financial institutions to evaluate the likelihood of client default, thereby facilitating the determination of credit limits and interest rates (Marini & Manfrin, 2020). Moreover, the use of alternative data such as consumption behavior and payment history has proven to be an effective tool for complementing traditional financial information and enhancing the accuracy of credit evaluations (Lassance & Ternoski, 2021).

Grunert et al. (2005) and Altman et al. (2010) emphasize the importance of incorporating qualitative variables, in addition to quantitative ones, in credit analysis models—especially for small and medium-sized enterprises. This integration provides a more comprehensive understanding of solvency and reduces default risks. Similarly, Khemakehem and Boujelbene (2017) highlight the use of not only financial variables but also non-financial ones, enabling a

more holistic analysis of consumers' credit profiles, particularly for those without formal financial histories, often referred to as "credit invisibles".

Credit commitment conditions are determined by assessing risks, weighing both favorable and unfavorable factors. These conditions are inherently linked to the micro- and macroeconomic environment, which enables institutions to extend credit through installment plans, deferred payments, checks, payment slips, store credit, or proprietary credit cards with greater confidence, thereby significantly reducing default risks (Lassance & Ternoski, 2021).

2.3 Use of Machine Learning in Credit Risk Analysis

According to Wanzeller et al. (2023), Big Data represents a vast collection of datasets that includes both structured and unstructured information from sources such as financial transactions, social media, and IoT sensors. In the context of credit granting, the use of Big Data enables the analysis of diverse information sources, encompassing both traditional credit histories and behavioral patterns (Jordan & Mitchell, 2015). The implementation of Big Data, Artificial Intelligence, Data Mining, and machine learning brings about a substantial transformation in this process, providing robust analytical capabilities (Timotio et al., 2024) and fostering the development of an efficient credit system that is well-adapted to the needs of both customers and the financial market (Wanzeller et al., 2023).

Credit scoring is primarily a classification problem, meaning that credit applicants must be assigned to a category based on their probability of default, as defined by the parameters of the Basel II Accord (Lessmann, 2015). Most of the classification methods employed are divided into two categories: statistical and machine learning (Dumitrescu et al., 2022). While the statistical approach aims to infer the relationships between attributes, the priority of machine learning is predictive performance (Montevechi et al., 2024), significantly impacting conventional classification methods by leveraging the processing of large volumes of data (Lessmann, 2015).

Dumitrescu et al. (2022) emphasize that, in the context of credit scoring, ensemble methods based on decision trees—such as random forest—demonstrate superior performance compared to models based on logistic regression. However, the authors also note that the latter remains the industry benchmark in credit risk management, primarily because the lack of interpretability of ensemble methods does not meet the requirements set forth by financial regulators (Dumitrescu et al., 2022).

According to Leo et al. (2019), machine learning possesses the ability to detect significant patterns in data, making it an invaluable tool for any task that demands the extraction of meaningful information from datasets. The authors argue that the growing adoption of machine learning has been driven by its potential to reduce costs, as well as enhance productivity and risk management.

Jordan and Mitchell (2015) point out that machine learning can identify complex patterns, resulting in more accurate predictions, and that the combined use of various types of data not only improves credit evaluation but also contributes to more informed decision-making—thereby promoting financial inclusion and reducing credit risk. Similarly, Lassance and Ternoski (2021) highlight that the adoption of these technologies in risk assessment represents a crucial advancement for financial institutions, as it enhances the accuracy of evaluations, optimizes processes, and personalizes customer service. Consequently, the credit granting system becomes more efficient and better adapted to the needs of the financial market (Marini & Manfrin, 2020).

As highlighted by Ambavat (2021), the adoption of advanced techniques such as machine learning allows for the incorporation of a wide range of non-financial variables, resulting in more robust and comprehensive risk models. Recent literature has emphasized the importance of integrating both financial and non-financial data to create a holistic view of consumers' credit profiles. Grunert et al. (2005) and Altman et al. (2010) underscore that including non-financial

(qualitative) variables alongside financial (quantitative) variables in credit analysis models for small and medium-sized enterprises enhances the accuracy of these models.

In the study by Khemakhem and Boujelbene (2018), within the context of corporate solvency, a comparison is made between the accuracy of artificial neural network (ANN) techniques and decision trees in a group of Tunisian companies, considering both financial and non-financial variables. The results obtained by the authors indicate that decision trees are more efficient than ANNs in terms of credit risk prediction when using balanced data. Additionally, the study contributes to the understanding of credit risk forecasting by revealing the variables and their relationship with the dependent variable, thereby assisting financial analysts in more accurately predicting credit risk (Khemakhem & Boujelbene, 2018).

Lee and Shin (2020) describe distinct approaches to machine learning: supervised, unsupervised, and reinforcement learning. According to the authors, in the case of supervised learning, the model is trained with labeled data, which aids in both categorization and prediction. In this context, classification facilitates the categorization of observations, while prediction underpins decision-making. Lassance and Ternoski (2021) further note that, in credit risk analysis, logistic regression is widely used to predict default, with supervised learning being applied in algorithms such as k-nearest neighbors, Naive Bayes, decision trees, random forests, and neural networks.

According to Zöller and Huber (2021), among supervised methods, decision trees stand out due to their ease of interpretation, although they may require adjustments, whereas artificial neural networks capture complex relationships among variables. Additionally, random forests enhance predictive accuracy by combining several decision trees. Lee and Shin (2020) point out that Gradient Boosting Decision Trees (GBDT) emerge as a powerful technique for improving accuracy when dealing with complex data. In another study by Zhang and Song (2022), which investigates credit evaluation for SMEs, a model based on the GBDT algorithm combined with a Convolutional Neural Network (CNN) and logistic regression (LR) is proposed. The experimental simulation was carried out on a sample of 14.366 SMEs, yielding results that demonstrate the GBDT-CNN-LR model outperforms individual statistical methods, as these traditional models are affected by the initial feature engineering process and do not exhibit the same efficiency as GBDT.

In view of this, Table 1 presents a summary of machine learning techniques classified by learning types.

Table 1

Machine Learning Techniques

Supervised Learning
Decision trees are a popular technique in which data is segmented based on input variables, creating a model that is simple to understand and interpret. However, decision trees can suffer from overfitting, which limits their ability to generalize (Zöller & Huber, 2021).
Artificial neural networks mimic the functioning of the human brain and are used to capture complex relationships between input and output variables. They are composed of layers of interconnected neurons capable of learning intricate patterns, making them especially useful when the data exhibits a non-linear structure.
Random forests, on the other hand, consist of an ensemble of decision trees trained on different subsets of data, thereby increasing accuracy and reducing overfitting. They combine the predictions of multiple trees to produce a more robust and precise forecast (Avelar et al., 2022).
Support Vector Machine (SVM) is a supervised learning method that seeks to find the optimal hyperplane which separates data classes with the widest possible margin. It is effective in high-dimensional spaces and is particularly useful when the number of dimensions exceeds the number of samples (Avelar et al., 2022).
Gradient Boosting Decision Trees (GBDT) is a boosting method within the ensemble learning category that combines multiple weak models—typically simple decision trees—in a sequential manner, correcting the errors of previous models. This method is powerful for enhancing predictive accuracy, especially when dealing with complex and large datasets (Avelar et al., 2022). In essence, GBDT is an advanced supervised learning technique that aggregates several weak models, usually decision trees, to create a strong overall model

Unsupervised Learning
The algorithm works with unlabeled data and attempts to identify inherent patterns or structures within the dataset. Techniques such as clustering and dimensionality reduction are commonly employed in this context, with algorithms like k-means and Principal Component Analysis (PCA) serving as typical examples.

Source: Zöller e Huber (2021) e Avelar et al. (2022).

As reported by O'Neil (2016), the evolution of machine learning holds transformative potential, especially in credit analysis, but it also introduces challenges such as overfitting, transparency, and ethical concerns. Ongoing research in these areas is crucial to ensure that models are both reliable and applicable to the financial context. Moreover, with the increasing complexity of the credit market, advanced techniques like machine learning have been integrated into traditional credit risk models to enhance predictions and identify risk patterns with greater precision (Ambavat, 2021).

3 METHODOLOGICAL PROCEDURES

The classification based on research objectives is descriptive, which is a powerful category for portraying characteristics, behaviors, or phenomena while seeking to establish relationships between variables (Gil, 2009).

The classification based on the nature of the research is qualitative, employing the case study method a choice justified by the fact that this approach provides a deeper understanding of the phenomenon in its entirety and complexity (Creswell & Poth, 2016). The case study is a widely recognized method when aiming for a profound and detailed understanding of a specific phenomenon, its particularities, and the intricate interrelationships among its categories (Godoy, 2006). According to Yin (2009), the case study offers an intensive analysis of a specific situation, focusing on its unique aspects with the predominantly descriptive purpose of illuminating the reality under investigation, thereby broadening the pathways to its comprehension.

According to Yin (2009), case studies offer the best approach for investigating research questions that seek to explain "how" and "why" a phenomenon occurs, allowing for a deeper and more comprehensive understanding of the topic in question. The "how" is descriptive in nature, as the analyzed processes yield accounts of the observed facts, and, as Godoy (2006. p. 128) suggests, one chooses a "unit of analysis to establish the boundaries of the researcher's interest." By delimiting the focus of the research, it is determined whether the study will be conducted on a single case or on multiple cases—the latter allowing for comparisons and more robust results (Yin, 2009). Flick (2004) indicates that "we should always start with a single case, studied in depth, before undertaking comparative analyses."

According to Stake (1995), a case is studied when it demonstrates special interest and when details of its interaction with the context are sought. A case study "is the study of the particularity and complexity of a single case, achieving its understanding within significant circumstances"(Stake, 1995. p. 11).

In a case study, the investigation begins with a specific unit of analysis, adopting predetermined criteria and multiple data sources—such as documents and archival records (usually quantitative) from the company and related agencies, interviews focused on the topics of the case study, direct observation (or when the observer participates as part of the context under study), or other physical artifacts (Yin, 2009. p. 102). Rather than merely delimiting a phenomenon, the case study proves more effective in broadening the understanding of it, allowing for an in-depth analysis of its nuances and complexities. According to Stake (2000), the ability to integrate quantitative data from diverse sources is a distinguishing feature of case studies compared to other qualitative methodologies, enabling a more comprehensive understanding of the phenomenon under investigation. Furthermore, Yazan (2016) notes that Robert Yin is a key reference in the

case study method, characterized by an objective, positivist perspective that seeks to address both qualitative and quantitative studies within the scope of the method.

The diversity of evidence sources seeks to mitigate the potential weaknesses raised by those who question the method's validity. Stake (1995) emphasizes the importance of triangulation for data validation. Denzin (1978) identified four basic types of triangulation:

> 1) Data Triangulation – the use of a variety of data sources in a study; 2) Investigator Triangulation – the involvement of several and diverse researchers in the investigation; 3) Theory Triangulation – the application of multiple perspectives to interpret a single set of data; e 4) Methodological Triangulation – the employment of multiple methods to study the same problem. According to Denzin (1978), triangulation can combine both qualitative and quantitative data collection methods and sources—such as interviews, questionnaires, observations and field notes, documents, among others).

According to Yin (2009. p. 49), the choice of a single case in case study research may be indicated under certain circumstances, "justified under five aspects: critical, unusual, typical, revelatory, or longitudinal." Thus, based on this methodological foundation, the company investigated in this research represents a case of special interest (Stake, 1995), requiring an in-depth single case study aimed at facilitating a rich analysis of the work. Furthermore, as noted by Yin (2009. p. 49), the case falls under the revelatory aspect "by allowing access to information that is not easily available and the possibility of disclosure," which is uncommon in the financial sector. Yin (2009) also cautions that although single case studies can be valuable, it is crucial to exercise caution when extrapolating their conclusions to other contexts, in order to avoid unfounded generalizations.

Supporting Stake (1995), the unit of analysis in this research, Banco BS2. is of special interest also due to the possibility of gaining a deeper understanding of an AI-based credit risk management system specifically, machine learning which is inherently complex because of its technical nature. The aim is to analyze the impacts related to default effects, which would normally be very challenging due to difficulties in accessing the data.

3.1 The BS2 Bank

BS2 Bank (2024) was founded by the Pentagna Guimarães family as Bonsucesso Bank S.A. in the 1990s. Originating from Minas Gerais, the bank formed a joint venture with Santander Bank in 2015. establishing Olé Consignado Bank. In 2017, Bonsucesso Bank repositioned itself in the market by focusing on digital products and rebranded as BS2 BankS.A., with its headquarters in Belo Horizonte, Minas Gerais. As part of its digital transformation, the bank has developed a technological platform specifically focused on serving SMEs and corporate clients (Banco BS2. 2024).

The product offering of BS2 Bank is built on four pillars: (1) credit products, (2) foreign exchange solutions, (3) cash management, and (4) insurance, with an important portion of the business represented by services (Banco BS2. 2024). In April 2024. the bank received a double upgrade from Moody's Local (2024), moving from a BBB+ rating to an A rating, due to its results and capitalization levels. In December 2023. BS2 reported consolidated total assets of BRL 12.7 billion and equity of BRL 741 million. Compared to 2022. the bank recorded a 61% increase in net income in 2023. reaching BRL 85 million. Additionally, there was a significant increase in the credit portfolio, with the volume of foreign exchange growing by 35% and cash management transactions increasing by 26%.

3.2 Data Collection

Stake (1995) emphasizes the importance of triangulation for data validation. Empirical evidence was collected through data triangulation (Denzin, 1978) using observation, interviews,

and document analysis techniques within a specific method in this case, a case study to gather and interpret the data. The use of multiple sources of evidence aims to avoid the errors noted by critics of the method by incorporating an analysis of the company's internal documents (technical manuals, reports on adopted criteria, selected financial variables, etc.) related to the credit risk management system, direct and participatory observation (with the researcher being part of the reality under study), and interviews conducted during June 2024.

he interview data collection instrument was a semi-structured guide developed based on the research protocol and literature studies, with the objective of "giving the interviewer flexibility to order and formulate the questions during the interview" (Godoi, Bandeira-de-Melo & Silva, 2010. p. 304). It was applied to the Head of Decision Science & Analytics at the company. The aim was to capture the interviewee's perspective in depth, in order to obtain a detailed understanding of the practices and challenges involved in implementing machine learning models within the context of Banco BS2. The interviews were recorded and transcribed, and the data were subsequently analyzed using content analysis.

The quantitative data analysis was conducted after the interviews, utilizing historical credit data from BS2 Bank, which employs both statistical methods and machine learning techniques. The Bank authorized access to the data and documents through a consent form. In order to evaluate which technique to use, performance metrics such as Accuracy, F1 Score, Area Under the ROC Curve (AUC-ROC), Precision, and Recall were obtained. The algorithms tested included Neural Networks, Decision Trees, Random Forests, Gradient Boosting Decision Trees, and Logistic Regression. The objective was to identify the most suitable model for predicting credit risk. Thus, to select the algorithm that provided the best performance metrics enabling a comparative analysis of the adopted approaches it was necessary to run the selected sample for each model using the software employed in this study, Scikit-Learn (https://scikit-learn.org/stable/), a dedicated machine learning tool in Python. The Gradient Boosting Decision Trees (GBDT) algorithm was ultimately chosen, as can be seen in Section 5. which presents the analysis and results.

3.2.1 Sample Used for Algorithm Evaluation

The sample consists of data from 10.000 small and medium-sized companies with open accounts at BS2 Bank that are considered to be in good standing, along with 700 companies that have defaulted on credit. The credit risk evaluation for these companies is classified by annual revenue: companies with annual revenues between R$ 1.000.000 and R$ 10.000.000 are considered small, while those with revenues between R$ 10.000.000 and R$ 30.000.000 are classified as medium. The total clientele amounts to approximately 150.000 clients. Analysis of individual consumers (PF) is not performed, as the bank is focused on companies and, therefore, does not maintain a PF portfolio.

3.2.2 Interview Guide

The interview protocol was developed based on the studies referenced in Table 2 and, as can be observed, covered seven topics. It focused primarily on questions related to machine learning techniques, such as the selection and implementation of the algorithm, data preparation, performance metrics, the integration of the models into the bank's credit decision system, security, challenges, and future perspective.

Table 2

Interview Guide

Questions by Topic	Reference Articles
1 Context and Implementation of Machine Learning (Gradient Boosting Decision Trees - GBDT*)*	Friedman (2001)
1.1 What were the main reasons that led BS2 Bankto adopt the GBDT algorithm for credit analysis?	
1.2 What was the process for implementing the GBDT algorithm in the bank's credit system?	
1.3. What were the main challenges encountered during the implementation of GBDT?	
2. Variable Selection and Importance	Chen e Guestrin (2016)
2.1. What techniques were used by the GBDT algorithm to select the most important variables?	
2.2. How did the GBDT algorithm help in identifying the most relevant variables for the credit model?	
2.3. How did the inclusion of non-financial variables impact the accuracy and effectiveness of the credit model?	
3. GBDT Performance Indicators	Berrar (2019)
3.1. Which performance metrics (e.g., accuracy rate, recall, F1-score, AUC-ROC) were used to evaluate the model's effectiveness?	
3.2. What were the most significant results obtained through these metrics?	
3.3. How does the model handle data imbalance in the samples?	
4. Practical Benefits and Results	Lessmann et al. (2015)
4.1. What were the main benefits observed after the implementation of GBDT in credit analysis?	
4.2. How did the use of this algorithm improve the bank's ability to identify and mitigate credit risks?	
4.3. Can you provide specific examples of how real-time analysis affected the bank's credit policies?	
5. Feedback and Future Perspectives	Ribeiro, Singh e Guestrin (2016)
5.1. What was the feedback from credit analysts and other stakeholders regarding the use of GBDT?	
5.2. Are there plans to incorporate other AI and ML techniques in credit analysis in the future?	
5.3. What additional improvements or adjustments are planned to further optimize the credit model?	
6. Ethical and Regulatory Considerations	Wachter, Mittelstadt e Floridi, 2017
6.1. How does the bank ensure compliance with financial regulations when using AI and ML models?	
6.2. What measures have been adopted to guarantee transparency and ethics in the use of customer data?	
6.3. Were there any concerns regarding data privacy during the implementation of the model? If so, how were they addressed?	
7. Team Integration and Training	Davenport e Harris (2007)
7.1. How was the process of integrating the data team with the other areas of the bank?	
7.2. What training and capacity-building programs were offered to employees to enable them to effectively use and interpret the results of GBDT?	
7.3. What were the main challenges faced by the team during the transition to using this new model?	

3.3 Data Preparation

3.3.1 Variable Selection and Algorithm Performance Metrics

Variable selection is an essential step in developing a predictive model, as it ensures the quality and validity of the obtained results by incorporating the significance of each variable into the developed model (see item 5.1.2). The performance analysis of machine learning algorithms, especially in credit risk contexts, requires the use of specific metrics that assess the model's effectiveness in predicting critical events, such as default. Among the most commonly used metrics are accuracy, precision, recall, and F1-score, which, according to Provost and Fawcett (2001), are fundamental for identifying the robustness of classifiers in imbalanced data scenarios, a common occurrence in the financial sector. To ensure a detailed analysis, the present study applied the Scikit-learn software a widely recognized Python library known for its capability to facilitate the development and evaluation of machine learning models (Pedregosa et al., 2011) to the algorithms selected by Banco BS2. Each of the models was fine-tuned based on financial and non-financial variables previously identified as relevant in the paper, weighted according to their impact on risk prediction, thereby providing a comprehensive view of the performance of the tested algorithms.

The effectiveness of a classification model is traditionally measured by accuracy, which provides the proportion of borrowers correctly classified (true positives and true negatives) out of the total set of borrowers (true positives, true negatives, false positives, and false negatives), although this measure can be ineffective when the data is imbalanced (Khandani et al., 2010). The performance metrics evaluated included Accuracy, F1 Score, Area Under the ROC Curve (AUC-ROC), Precision, and Recall (or Sensitivity). The objective is to identify the most suitable model for predicting credit risk.

Studies on the use of accuracy in imbalanced data situations have sparked debates (Provost & Fawcett, 2001; Sun et al., 2007). From the confusion matrix, performance can be obtained (see Table 3). Thus, the elements along the main diagonal represent the correct decisions the number of true negatives (TN) and true positives (TP) while the elements outside this diagonal represent the errors made, namely, the number of false positives (FP) and false negatives (FN) (Castro & Braga, 2011).

Table 3
Confusion Matrix for a Binary Classifier.

	Prediction (y = 0)	Prediction (y = 1)
real (y = 0)	**VN**	FP
real (y = 1)	FN	**VP**

Luque et al. (2019) emphasize that essential metrics can be extracted from the confusion matrix, as described in Table 4.

Table 4
Model Performance Measures

Accuracy or Rate of Correctly Classified Cases	Indicates the overall performance of the model by measuring the proportion of cases that the model correctly classified among all classifications. $$Acurácia = \frac{VP + VN}{VP + VN + FP + FN}$$

recision or Rate of Predicted Positives:	Indicates, among all positive class predictions made by the model, how many are correct. $Precisão = \frac{VP}{VP + FP}$
Recall/Sensitivity/Revocation or True Positive Rate	Indicates, among all cases where the positive class is expected, how many were correctly identified. $Recall = \frac{VP}{VP + FN}$
Specificity	Measures the proportion of negative cases that are correctly classified as negative. $Specificity = \frac{VN}{VN + FP}$
F1-Score	The harmonic mean between precision and recall. $F1\ Score = 2 * \frac{Precisão * Recall}{Precisão + Recall}$

Furthermore, the ROC Curve (Receiver Operating Characteristic) is also derived from the Confusion Matrix, generated on a graphical plane and representing the trade-off between Sensitivity and Specificity indicators. The ROC Curve provides an estimate of the classifier's discriminative ability in terms of error probability (Castro & Braga, 2011). The AUC – Area Under the Curve offers a metric to evaluate, on average, which algorithm performs best; the larger the area between the ROC Curve and the main diagonal, the better the model's performance.

According to Botelho and Tostes (2011), the Kolmogorov-Smirnov (KS) test is a non-parametric statistic designed to test whether the distributions of two groups are equal. In classification problems, it is frequently used to measure the ability of a classification model to distinguish between two classes. Furthermore, the maximum KS value indicates the cutoff point at which the model best differentiates between the two classes (Botelho & Tostes, 2011).

4 PRESENTATION AND ANALYSIS OF THE RESULTS

In this section, the results derived from the triangulation of the data collected from interview narratives, direct and participatory observation with corresponding field notes, and gathered documents are presented, providing sufficient elements to foster a comprehensive joint analysis.

4.1. Data Preparation

4.1.1. Choice of the Gradient Boosting Decision Trees (GBDT) Algorithm

Table 5 presents the results of the performance metrics for each model, enabling a comparative analysis among the following approaches: neural networks, decision tree, random forest, GBDT, and logistic regression. Although GBDT achieved the best results, the differences observed with the applied metrics did not yield statistically significant differences when compared to the other algorithms.

Table 5

Statistical Results of the Algorithms

Algorithms	Accuracy	Precision	*Recall*	F1 - Score
Neural Networks	0.95	0.85	0.79	0.78
Decision Tree	0.94	0.84	0.76	0.75
Floresta aleatória	0.96	0.90	0.80	0.81
GBDT	**0.97**	**0.92**	**0.83**	**0.83**
Logistic Regression	0.95	0.85	0.79	0.78

As elucidated by the interviewee regarding the context and selection of Machine Learning (Gradient Boosting Decision Trees - GBDT*):*

> After testing the neural networks, decision tree, random forest, GBDT, and logistic regression algorithms, GBDT yielded the best results in terms of metrics. Additionally, GBDT is known for its high precision and robustness, particularly in classification and regression tasks. It combines multiple weak models, usually decision trees, to create a strong model, resulting in superior predictive performance. Furthermore, the interpretability of decision trees facilitates understanding the factors that influence credit decisions, which is crucial for regulatory compliance and transparency.

In addition to the obtained results, the GBDT algorithm, as noted by Sun et al. (2007), was also chosen because it is renowned for its ability, when testing various datasets, to predict defaults, mitigate risks, and provide precise measurements of metrics such as accuracy, recall, F1-score, and the area under the ROC curve (Receiver Operating Characteristic). To clarify, the boosting method employs *n* decision trees with random samples; however, these trees are not independent since the learning process is sequential. Each tree is trained to minimize the error made by the previous trees. An extension of boosting algorithms is GBDT, where the error is minimized by using the residual errors from the previous trees. In other words, GBDT is based on training several decision trees sequentially, with each new tree aiming to correct the errors of its predecessors. Another factor supporting the choice of GBDT is that, according to Lee and Shin (2020), Gradient Boosting Decision Trees (GBDT) emerge as a powerful technique for enhancing accuracy in complex data.

To run the credit analysis modeling using the GBDT algorithm, BS2 Bank selected approximately 10,000 samples of compliant clients and 700 defaulted clients. The bank considers a client to be in default if they have at least one account receivable with an outstanding principal of R$100 and if the receivables are overdue by 60 days after the due date. Regarding the context and implementation of GBDT at BS2 Bank, the interviewee stated that it involved six steps:

> 1 Conducting a requirements analysis to identify the specific needs of the credit model; 2. Data cleaning and preprocessing to ensure data quality and consistency; 3. raining: using a set of historical credit data to train the algorithm; 4. Applying techniques such as cross-validation to avoid overfitting; 5. Validating and testing the model to evaluate its performance; and 6. Integrating the model into the bank's credit system, with continuous monitoring for adjustments and improvements.

According to the interviewee, several challenges were encountered during the implementation of GBDT. These included the need to handle large volumes of data and manage the multiple data sources that BS2 Bank accesses for credit analysis. Another significant challenge was the class imbalance in the dataset, with the majority of records corresponding to non-defaulting clients. Additionally, there was a need to fine-tune the model's hyperparameters to optimize its performance, a process that required considerable time and computational resources.

> Regarding data imbalance, the interviewee emphasized that the model employed resampling techniques such as oversampling the minority classes or undersampling the majority classes to address the imbalance in the data. Additionally, class weight adjustment techniques were applied during model training to assign greater importance to the minority classes. These approaches helped ensure that the model was not biased in favor of the majority class, thereby improving its ability to accurately predict instances of the minority class.

4.1.2 The Selection and Evaluation of Variable Importance by GBDT

The selection of variables for credit analysis at Banco BS2 was conducted rigorously by accessing multiple data sources: four credit bureaus, internal information, public information, and reputational web research. This comprehensive approach was essential for capturing a complete

overview for credit analysis and for presenting the findings in a technical and scientific manner. After data collection, an extensive set of variables was initially considered for the analysis. Following a process of cleaning and addressing missing or incomplete data, the samples were then selected. The process involved using intrinsic techniques of the algorithm to determine the importance of each variable, ensuring that only the most relevant ones were included in the final model.

Regarding the selection and importance of variables, the interviewee stated:

> The GBDT algorithm employs an iterative process to select significant variables based on their contribution to reducing the model's error. During training, the algorithm calculates the importance of each variable based on the improvement it provides to the splits in each decision tree. The variable importance metric can be analyzed by summing the reductions in splitting criteria (such as Gini or entropy) for each variable across all the trees in the forest. These importances are then normalized, providing a relative measure of each variable's relevance within the model
>
> .

In other words, for greater technical clarity, during training, the algorithm creates several decision trees, where each tree corrects the errors of the previous ones. At each decision node, the algorithm selects the variable that most reduces the predictive error. The importance of each variable is then measured by summing the reductions in the splitting criteria using entropy across all trees. The significance of the variables is calculated based on the improvement each variable provides to the splits in each decision tree. This is achieved by summing the error reduction provided by each variable throughout all the trees in the model. Variables with the highest sums are considered the most important. This process quickly identifies which variables have the greatest impact on predicting credit risk, resulting in a more efficient and accurate model.

Variable selection is an essential step in conducting robust quantitative research, as it ensures the quality and validity of the results obtained. To identify the most relevant and significant variables for the GBDT algorithm model, both financial and non-financial variables were selected based on the significance extracted from multiple regression, as listed in Table 6.

Regarding the impact of including non-financial variables on the accuracy and effectiveness of the credit model, the interviewee emphasized:

> The inclusion of non-financial variables increased the model's accuracy and effectiveness by providing a more holistic view of the customers' credit profiles. These variables added an extra layer of information that complements traditional financial data, enabling the model to capture behavioral and socioeconomic aspects of the customers. As a result, the model became more robust and capable of predicting default risk with greater accuracy, especially in cases where financial variables alone were insufficient for a complete evaluation.

This finding from the interviewee aligns with studies such as Ambavat (2021), Grunert et al. (2005), and Altman et al. (2010), which confirm that the adoption of machine learning techniques enables the incorporation of a wide range of non-financial variables, resulting in more robust and comprehensive risk models that improve their accuracy.

Table 6

Financial and non-financial variables used in the model with their respective significances extracted by Scikit-Learn

Variable	Signif. (%)	Description
general_lost_pct_risco	32.27	Represents the overall percentage risk of default (overdue + loss). It is a direct risk metric that reflects the probability of default, making it crucial for predicting credit behavior.

years_since_first_relationship (non-financial)	8.84	Represents the time since the beginning of the banking relationship. Customers with longer relationships tend to have a more reliable, solid, and predictable history, reducing uncertainty in credit analysi.
chk_esp_cred_rot_a_vencer_pct_ risco_max_prev_10m	8.73	Indicates the maximum amount of overdraft and revolving credit due in the next 10 months. High amounts may indicate liquidity problems and an increased risk of default.
cnt_declined_reasons (non-financial)	7,21	Indicates the number of credit refusals, where a high number may signal recurring credit issues, as multiple refusals are associated with a higher risk of default.
general_limite_do_cred_360m_sum	5.47	Refers to the sum of credit limits up to 365 days obtained from other institutions. This index reflects the ability to access credit, influencing both liquidity and financial risk. Elevated credit limits may indicate a compromised financial capacity regarding risk exposure.
days_since_oldest_board_member (non-financial)	5.24	Measures the number of days since the last update of the financial status of board members. Companies with regular updates tend to be more transparent, demonstrating proper management, diligent governance, and strong internal controls, which ultimately reduce the risk of default.
v001r_social_capital	5.04	This indicator refers to the company's share capital. A high share capital can indicate a solid financial base, suggesting that the company has a greater capacity to withstand financial adversities and, consequently, a reduced risk of default.
chk_esp_cred_rot_a_vencer_pct_risco	5.02	Overdraft + Revolving Credit as a Percentage of Total Risk.
min_lj_ij3_prev_6m	2.84	Minimum Long-Term Interest Rate for Loans Over 360 Days (Last 10 Months) The minimum long-term interest rate is estimated for loans exceeding 360 days. Lower interest rates may indicate more favorable credit conditions, reflecting the company's ability to access credit at lower costs.
general_limite_do_cred_sum_pct_risco	2.62	Refers to the total credit limits obtained from other institutions for up to 365 days. This index reflects the company's ability to access credit, influencing both liquidity and financial risk. High credit limits may indicate a compromised financial capacity and increased exposure to risk.
v254r_cnt_board_members (não financeira)	2.42	Refers to the total number of board members. A larger board may indicate a more robust governance structure. Strong governance is associated with a lower risk of default due to improved decision-making, better oversight, and mitigation of agency risks.
avg_final_number_prediction_prev_6m	2.38	Average Forecast Over the Last 10 Months.
days_since_stopped_simples	2.35	Days Since SIMPLES Status Was Interrupted.
general_limite_de_cred_sum	2.17	Total Credit Limit.
general_limite_trend_cred_sum	1.90	Credit Limit Growth Over the Last 10 Months.
v204r_is_popular_domain	1.80	Indicator for Popular Email Domain Usage (Gmail, Yahoo)
cnt_late_not_to_calc_interest	1.49	Count of Irregular Installments Over the Last 10 Months.
special_interest	1.37	Average Interest Rate Across All Loan Types.
primary_median_ij3_prev_6m	1.01	Interest Rate on the Highest Exposure in the Last 10 Months.

4.2 Training / Cross-Validation / Testing of GBDT

A clear definition of default helps ensure consistency in data labeling. **Table 7** describes the data used to train the credit analysis model, highlighting the distribution of transactions and defaults. Thus, the data was divided into training, cross-validation (CV), and testing subsets to evaluate the model at different development stages, ensuring its robustness and generalization capability. Each subset contains information on the number of transactions and the default rate. The reduction in the number of transactions from training to cross-validation and testing reflects common data-splitting practices in machine learning. During preprocessing, some transactions may be filtered or excluded due to missing data, inconsistencies, or outliers. This cleaning process can reduce the total number of transactions available for validation and testing.

Table 7
Data Split for Training / Testing

2020 - 09	Transactions	*Defaults*	Default Rate
Training Data	5469	374	6.8%
Cross-Validation (CV)	2344	160	6.8%
Test Data	1908	161	8.4%

To ensure that the class distribution (e.g., default rate) remains representative in each subset, stratification can be used. This technique ensures that the proportion of defaults is maintained consistently across the training, validation, and test sets, which may result in the exclusion of some transactions to preserve this balance. The training set contains more data than the combined validation and test sets. This is a common practice, as training requires a larger dataset to allow the model to learn robust patterns and improve its predictive performance.

According to Table 7, the key data analysis points are as follows: 1. Consistent Default Rate in Training and Validation: The default rate in the training and cross-validation sets is exactly the same (6.8%), which is beneficial for ensuring that the model is trained on data with similar characteristics; 2. Higher Default Rate in the Test Data: The default rate in the test set (8.4%) is higher than in the other subsets. This may indicate that the test set represents a different period or a higher-risk segment, implying greater variability or difficulty in the unseen data. In a way, this is expected and desirable for a robust evaluation of the model. However, it may also signal a limitation in the model's ability to generalize perfectly to new data.

4.3 Score Systems Used by BS2

BS2 employs two scoring models for analyzing eligible samples: Application Score – Based on data provided by the Serasa credit bureau. The Application and Behavioral scoring systems support the model for continuous monitoring and credit line adjustments throughout the duration of the customer's credit agreement.

The Application Score (Table 8) is used in the pre-screening process to identify customers with potential delays exceeding 60 days, based on data provided by the Serasa credit bureau. The Behavioral Score (Table 8) is used to identify customers with a tendency to accumulate significant delaysmdefined as having 10% or more past-due payments across all banking credit lines. This analysis focuses on data from the Central Bank of Brazil's Credit Information System (SCR). This score is used for customer monitoring, real-time decision-making, and final credit assessment. The evaluation of both scores is based on four key indicators: F1-score, Kolmogorov-Smirnov (KS) Indicator, KS-Analog Indicator, and Area Under the ROC Curve (AUC-ROC). The system evaluation relies on the F1-score for training, validation, and monthly performance monitoring, and the model also undergoes external validation.

Table 8
Metrics of Application and Behavioral Scores Extracted by the Model

Score Type	Application	Pontuation	Behavioral
Metric	**Value**	**Metric**	**Value**
F1 Score	0.66	F1 Score	0.77
KS	0.783	KS	0.691
KS Analog	0.766	KS Analog	0.742
AUC ROC	0.94	AUC ROC	0.919

Based on Table 8, the explanation of the metrics extracted from the Application Score model is:

- F1 Score: A performance metric that combines precision and recall, indicating a balance between these two measures (0.66). The closer to 1, the higher the precision;
- KS Indicator (Kolmogorov-Smirnov): Measures the maximum difference between the cumulative distributions of two populations, in this case, good and bad payers. The closer to 1, the smaller the difference between the classes (0.783);
- KS Analog: Analogous to KS, used for internal validations (0.766). In addition to validating KS, it helps reduce differences between classes;

AUC-ROC: The Area Under the ROC Curve (Receiver Operating Characteristic) measures the model's ability to distinguish between positive classes (true positives) and false positives. Values close to 1.0 indicate high model effectiveness. The ROC curve and AUC are used to evaluate the model's ability to discriminate between classes. An AUC-ROC of 0.94 indicates excellent discrimination capability.

Table 9 presents the metrics and the most important features used by the Application Score model, along with their respective importance values and cumulative scores. The model includes performance metrics such as precision, recall (or sensitivity), F1-score, and the number of occurrences for the "False" and "True" classes. Precision is very high for the "False" class (0.98), meaning the model is highly effective at correctly identifying non-defaulting cases. The recall (or sensitivity) for the "True" class (0.71) indicates that the model is also effective in identifying defaulting cases).

Table 9
Application Model Metrics

Classification	Precision	*Recall*	F1-Score
False	0.98	0.97	0.97
True	0.61	0.71	0.66
Macro Avg	0.80	0.84	0.82

Regarding the Application Score modeling results for performance metrics—precision, recall (or sensitivity), F1-score, and the number of occurrences for the "False" and "True" classes as shown in Table 9, it can be observed that:

- Precision is very high for the "False" class (0.98), meaning the model is excellent at correctly identifying non-defaulting cases.
- Recall (or Sensitivity) for the "True" class (0.71) indicates that the model is also effective in identifying defaulting cases.
- F1-Score Macro Average is used to evaluate the model's ability to distinguish between classes. The macro average result of 0.82 indicates a high discrimination capability.

4.3.1 Behavioral Score Evaluation (BS2)

The Behavioral Score enables continuous monitoring of customers throughout their relationship with the bank, unlike the Application Score, which is based solely on initial information. This allows the model to detect changes in risk profile over time.

Additionally, the Behavioral Model incorporates updated customer data, such as account transactions, revolving credit usage, and payment history. These data points provide a more comprehensive and accurate reflection of the customer's current financial situation (Table 8) is:

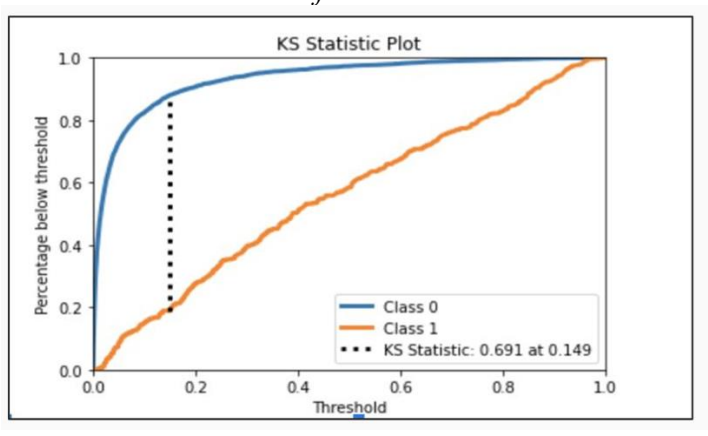
- F1 Score: 0.77 – A higher F1-score compared to the Application Score, indicating a better balance between precision and recall for this type of scoring.
- KS (Kolmogorov-Smirnov): 0.691 – Still a good metric, though slightly lower than the KS value of the Application Score.
- KS Analog: 0.742 – Similar to KS, used for internal validations.
- AUC-ROC: 0.919 – Indicates high effectiveness in distinguishing between good and bad financial behaviors, although slightly lower than the AUC-ROC of the Application Score.

The Behavioral Score is based on the customer's behavioral history, including payments, defaults, and financial transactions, whereas the Application Score focuses on the information provided at the time of the credit application. BS2 opted for the Behavioral Model, which demonstrated a higher F1-score (0.77), indicating that it is more accurate in correctly predicting both compliant and defaulting customers.

To compare the cumulative distribution of scores between two classes, BS2 uses the Kolmogorov-Smirnov (KS) statistic (Figure 1), where: "**"Class 0"** likely represents good payers and "Class 1" represents bad payers. The blue curve represents the cumulative distribution of Class 0, while the orange curve represents the cumulative distribution of Class 1 (Figure 1). The dashed vertical line indicates the maximum difference between the two distributions, which corresponds to the KS statistic. As shown in Figure 1, the KS statistic is 0.691, indicating a significant difference between the distributions of the two classes. This suggests that the model effectively distinguishes between good and bad payers, as there is a clear separation between the score distributions of both classes.

Figure 1

KS Metric Extracted from the Behavioral Model



The green vertical line in Figure 2 represents the maximum KS Analog, which is 0.742. Similar to the traditional KS graph, the KS Analog shows the maximum difference between the two distributions. The high value of 0.742 suggests a strong separation between defaulting and non-defaulting customers, indicating that the model is effective in distinguishing these two states.

Figure 2
KS Analog Metric Extracted from the Behavioral Model

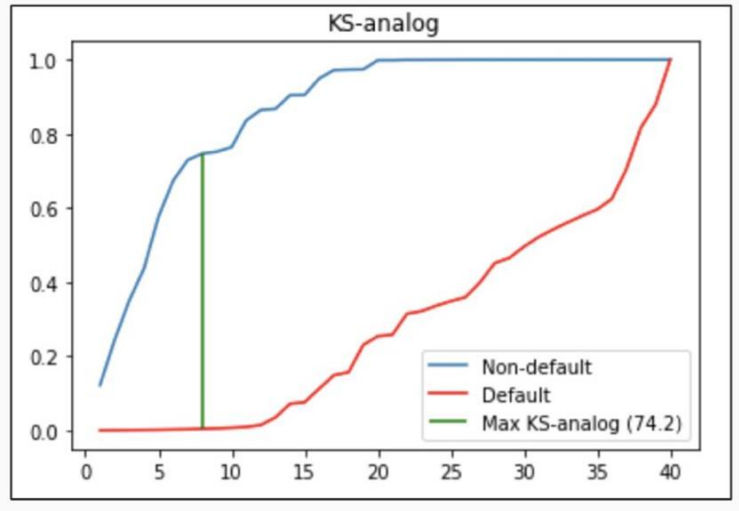
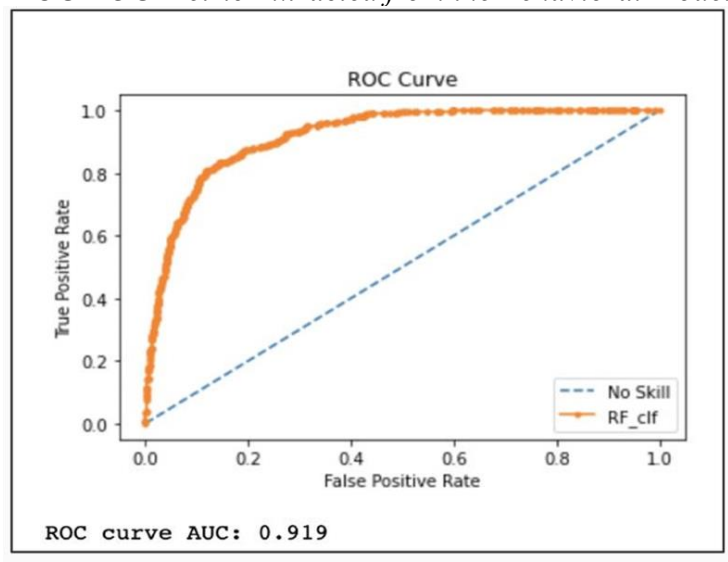


Figure 3 describes the performance of the classification model by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) for different decision thresholds. The AUC of 0.919 indicates that the model has excellent discriminatory power between the classes (good and bad payers). The closer the AUC value is to 1, the better the model's performance.

Figure 3
AUC-ROC Metric Extracted from the Behavioral Model



In summary, the results in Table 10 indicate that, although the Behavioral Score exhibits lower KS and AUC-ROC indices compared to the Application Score, the credit model still demonstrates strong performance in distinguishing between good and bad payers. This is reflected in the high separation between score distributions (high KS and KS Analog statistics) as well as its good discriminatory ability, as indicated by the AUC-ROC curve.

Table 10
Metrics of Application and Behavioral Scores at BS2

	KS	KS Analog	F1	AUC (Area Under ROC Curve)
Application	0.783	0.766	0.66	0.94
Behavioral	0.691	0.742	0.77	0.919

Regarding the most significant results obtained through the metrics, the interviewee highlighted that:

> When comparing the models, even though the Behavioral Score presented lower indices than the Application Score, the results showed that the GBDT model achieved high precision and recall, indicating its effectiveness in identifying both compliant and defaulting clients. The F1-score provided a balanced measure of the model's performance, demonstrating a good trade-off between precision and recall. Additionally, the AUC-ROC was high, suggesting that the model has strong discriminatory power between classes. These significant results indicate that the model is robust and reliable for credit decision-making.

These results align with the literature, as evidenced by Zhou et al. (2019), who highlight that the F1-score is a common criterion in information retrieval for evaluating classification model performance. Both this metric and the AUC-ROC, which also demonstrated a high value, reinforce the model's effectiveness. The preference for these metrics is due to the fact that, although accuracy is widely used in many studies to assess model performance, Khemakhem and Boujelbene (2018) point out that accuracy can lead to biased results when dealing with imbalanced data, potentially leading to the misselection of a predictive model.

4.4 BS2 Bank Default According to BCB Rule 2682

Resolution No. 2682 of the BCB establishes guidelines for credit risk classification and the corresponding provisioning for associated losses (BCB, 1999). This measure aims to ensure that financial institutions maintain an adequate capital reserve to cover potential losses, thereby promoting financial system stability.Risk classification is divided into eight levels, from A to H, with each level representing an increasing risk of default. Category A is considered low risk, requiring a provision of only 0.5%, while Category H, representing the highest risk, requires a full 100% provision.

Financial institutions are required to periodically review their credit portfolios and adjust provisions as needed, based on the reassessment of borrowers' default risk. This practice is essential to ensure that institutional balance sheets accurately reflect credit risk and are prepared to absorb potential losses. Table 11 presents BS2 Bank's default data from Dec/21 to Jun/24, as shown in Figure 4, which demonstrates a decline in the default rate after Jun/23, six months after the adoption of the new credit analysis model using the GBDT algorithm. This highlights the positive impact of the predictive credit model.

.

Table 11
BS2 Bank Default Rate (Dec/21 – Jun/24)

Default Rate	Dec/21	Jun/22	Dec/22	Jun/23	Dec/23	Jun/24
Over 30	0.29%	0.94%	2.44%	3.68%	2.21%	1.58%
Over 60	0.25%	0.50%	2.26%	3.45%	1.82%	1.21%
Over 90	0.17%	0.39%	0.40%	3.25%	1.75%	1.17%

Figure 4
BS2 Bank Default Rate (Dec/21 – Jun/24)

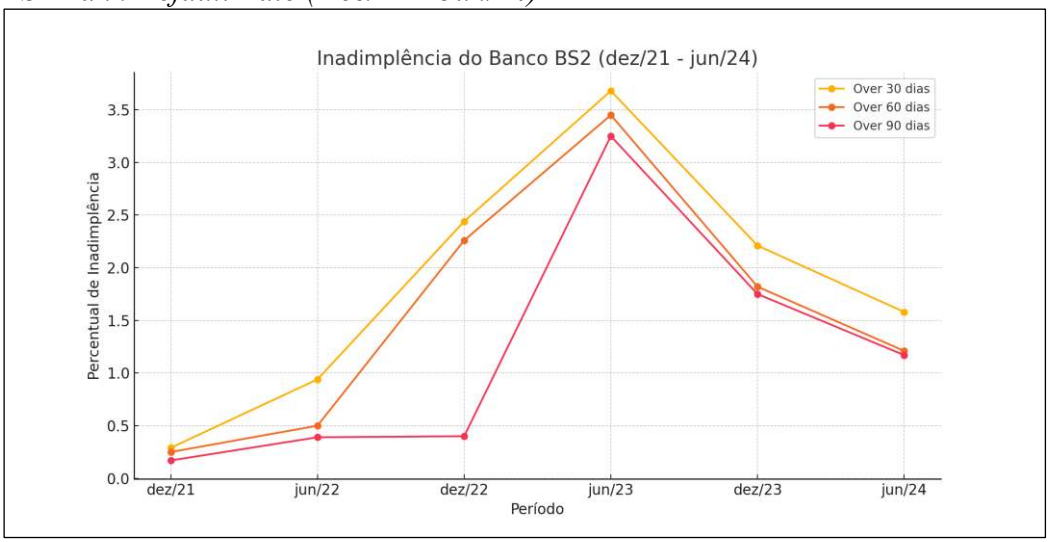
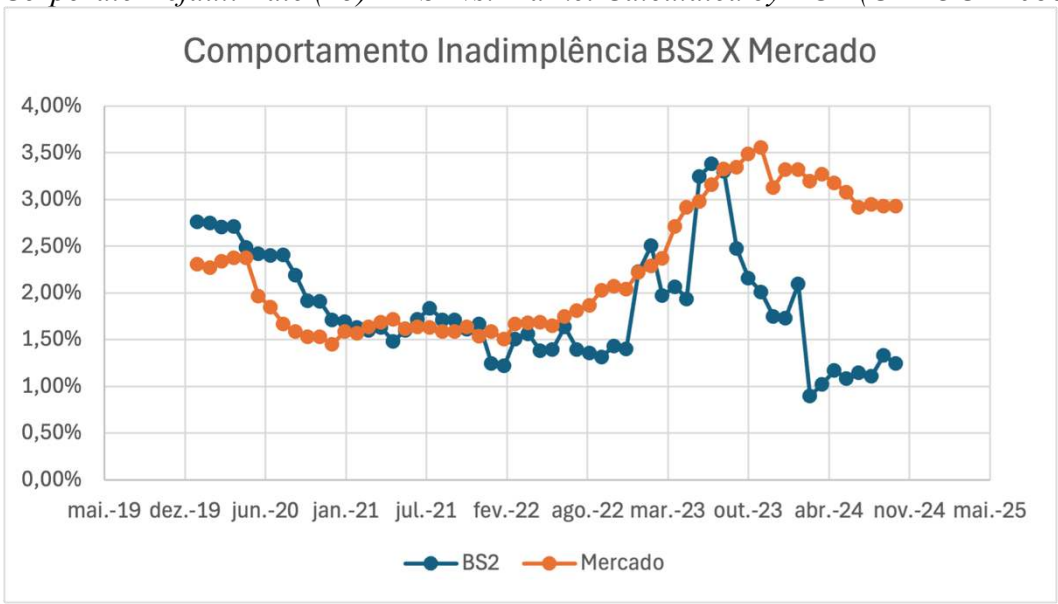


Figure 5 presents the BS2 Bank's corporate (PJ) default rate compared to the market default rate, as published in CADOC-21086 of the BCB (BCB, 2024). CADOC (*Document Catalog*) is a set of mandatory reports that financial institutions must submit to the Central Bank of Brazil (BCB).

Figure 5
Corporate Default Rate (PJ) – BS2 vs. Market Calculated by BCB (CADOC 21086)



Source: BCB-CADOC (2024)

The historical records of the default rate indicate that BS2's indices have aligned with market trends, showing a declining trend since July 2023. The default rate can be divided into three distinct phases: (1) 2020 – Pandemic, when default rates were above market levels due to economic instability; (2) 2021-2022 – Convergence with Market Rate, during which the default rate aligned with the market trend; and (3) 2023-2024 – Declining Rate, where a steady decrease in default rates followed the implementation of the new credit analysis model. This data reflects a significant

improvement in risk identification and credit decision-making efficiency after July 2023, marking six months since the adoption of the GBDT-based credit analysis model. However, to definitively confirm that these improvements in default indices were solely due to the new model, a longer observation period would be necessary.

4.5. Benefits of Implementing the GBDT-Based Model

> Regarding this topic, the interviewee emphasized that:
> The main benefits were those that led to greater accuracy in credit risk assessment, enabling a reduction in default rates over 60 and 90 days since its implementation in December 2022, and consequently, a decrease in financial losses. Furthermore, in the interviewee's words"The ability to identify credit risks early allowed for the implementation of more effective preventive measures and faster, data-driven credit decisions."

In other words, the interviewee emphasized that with the adoption of GBDT, the bank's ability to identify complex patterns and predict risk behaviors was significantly improved. By enabling early detection of defaults, preventive actions such as credit limit adjustments, debt renegotiation offers, or financial education programs can be effectively implemented.

The bank can quickly adjust its credit policies in response to economic fluctuations or changes in customer behavior. This was particularly evident during the pandemic period".

4.6 *Credit Analysts' Feedback / Future Perspectives*

The interviewee emphasized that the feedback from credit analysts was overwhelmingly positive, highlighting improvements in prediction accuracy and the ease of interpreting results, with "recognition of the positive impact on operational efficiency and the reduction of financial losses." Regarding future improvements or adjustments, "there are plans to further optimize the credit model by integrating new data sources, both internal and external, to enrich the dataset used by the model." This may include alternative data, such as rental payment history and utility bills, which can provide additional insights into customer credit behavio.

4.7 Ethical and Regulatory Considerations

According to the interviewee, BS2 Bank adopts a rigorous approach to ensuring compliance with financial regulations when using AI and machine learning models. This includes "the implementation of robust data governance policies, the conduction of regular compliance audits, and the maintenance of a dedicated compliance team." The models are continuously monitored and evaluated to ensure they comply with regulatory guidelines, such as the General Data Protection Law (LGPD) and other relevant international standards. To ensure privacy, BS2 has implemented multiple layers of data protection, including data encryption at rest and in transit, strict access control policies, and continuous monitoring of suspicious activities. Additionally, a data privacy committee was established to review and approve the use of data in the development of machine learning models, "ensuring that all practices align with current privacy regulations".

5 CONCLUSION

This case study aimed to analyze the use of a machine learning tool, specifically the Gradient Boosting Decision Tree (GBDT) algorithm, in credit risk management at Banco BS2. GBDT is an iterative decision tree algorithm that consists of multiple decision trees, where the final decision is based on combined conclusions from all trees (Zhang et al., 2018). Data triangulation was the methodological approach adopted to reduce noise and improve understanding, given the complexity of the topic. The implementation of the predictive model at the bank through the GBDT algorithm represents a significant advancement in identifying good payers, allowing for better predictive credit

profile assessment. The goal is to enhance credit management, reduce default rates, and optimize financial operations, directly impacting results by improving provisioning levels and default management. Additionally, the use of GBDT in credit analysis enables BS2 Bank to quickly adapt to changes in customer behavior and market conditions, ensuring greater flexibility and dynamism in its credit policies.

BS2 chose this algorithm after analyzing the performance metrics of various classification algorithms, including neural networks, decision trees, random forests, GBDT, and logistic regression. This allowed for a comparative analysis of the adopted approaches, considering their specific advantages, such as simplicity, ease of interpretation, and the ability to detect complex patterns. Additionally, as reported by the Head of Decision Science & Analytics, after conducting tests, this algorithm stood out in its ability to handle large volumes of data, address sample imbalance issues, and improve the selection of financial and non-financial variables based on their direct relevance to credit risk assessment. GBDT is well known for its high accuracy and robustness, particularly in classification and regression taskS.

Credit granting plays a crucial role in the financial health of institutions. According to Modigliani and Brumberg's (1954) life-cycle theory, an individual's payment capacity varies throughout their life. Young individuals tend to take on more credit to finance consumption and investments, while middle-aged individuals are more likely to save, and retirees typically consume their savings. Accurate risk assessment is a key factor in preventing defaults and financial losses.

The application of technologies has led to more precise and robust data analysis, while also effectively mitigating risks associated with credit granting, as suggested by Hand and Henley (1997). These modeling techniques have revolutionized how financial institutions assess credit risk. The integration of these technologies can transform data analysis and decision-making, resulting in a more efficient, accurate, and secure credit system. The ability to predict default behavior with greater accuracy and optimize operational processes are just some of the many advantages offered by artificial intelligence (AI) and machine learning.

The implementation of the GBDT algorithm, combined with credit bureau consultations in BS2 Bank's credit risk management, enabled the institution to develop a predictive model that improved operational efficiency, allowing for faster, data-driven credit decisions. Additionally, BS2 Bank's default rate, as recorded in the BCB-CADOC, has been declining since the implementation of the new credit analysis model based on the GBDT algorithm. In traditional credit assessment systems, sociodemographic data and loan application details are used as input variables for credit analysis models. However, the dynamic transaction history of applicants, which is a key metric for assessing repayment behavior, is typically not included in traditional credit evaluation systems.

Thus, the Behavioral Model introduced at BS2 aims to address this issue, aligning with the study by Zhang et al. (2018), which proposed a comprehensive evaluation method incorporating traditional data, individual sociodemographic information, loan application details, and dynamic behavioral transaction data from applicants. The test results from this particular study demonstrated that this approach significantly improved predictive performance, based on the most commonly used model evaluation criteria (Zhang et al., 2018). This is also reflected in the F1-score metric, which BS2 Bank uses as a benchmark to demonstrate model effectiveness and accuracy. Compared to Serasa's Application Score model, BS2's Behavioral Model achieved a higher F1-score of 0.77. The F1-score is a widely used metric in information retrieval to assess classification model performance (Zhou et al., 2019). Additionally, this study involves imbalanced data, where compliant companies dominate, making the best model the one that achieves the highest F1-score. The F1-score represents the weighted average of sensitivity (recall) and precision. Many studies also use accuracy to evaluate model performance; however, as highlighted

by Khemakhem & Boujelbene (2018), the issue with accuracy is that a high precision rate in imbalanced datasets can lead to biased results and ultimately poor selection of a predictive model.

To enhance the performance of the Behavioral Model, as emphasized by the Head of Decision Science & Analytics, BS2 may implement the following strategies:

- Enhancing Input Data by introducing LLMs (Large Language Models) into the credit analysis model to bring improvements: incorporating alternative data beyond financial behavioral data, including additional information such as social media data or online feedback (with customer consent), history of customer support interactions (emails, chat messages, call transcripts), consumer behavior analysis, and card usage patterns;
- Model Training: Using a larger historical dataset to train the model can enhance its predictive capability, with frequent model updates that is, regularly retraining the model to capture changes in customer behavior and economic conditions;
- Score Combination: Creating a hybrid model that combines the Application Score and the Behavioral Score can improve overall results. For example, the Application Model could be used for pre-selection, while the Behavioral Model would be applied in a second stage, focusing on clients approved in the initial screening.

Based on the findings of this study, future research should focus on the following topics:

- Broader application of ensemble methods, such as boosting techniques, including GBDT, for credit risk prediction, considering that ensemble methods have been underutilized in these models, as previously mentioned in this study.
- Future research should compare algorithms within these techniques to assess model performance over time.
- When collecting data samples for future studies, it is recommended to focus on homogeneous groups of defaulting and non-defaulting clients.
- A final recommendation is variable selection, as an optimal combination and selection methods can enhance the predictive model, allowing a performance comparison between individual algorithms and various combinations in the development of predictive models.

This study aims to contribute to the knowledge gained from Banco BS2's experience, which can serve as a model for other financial institutions seeking to improve their risk management practices by implementing a predictive model through machine learning. Additionally, given the social and economic consequences of credit risk management, the development of a more precise predictive model can enhance decision-making accuracy, becoming a strategic differentiator by fostering innovation. Finally, the knowledge shared through the BS2 Bank case study aims to support the exploration of other algorithms that may offer even better performance, demonstrating how the application of these technologies can contribute to a more resilient and stable financial system, with an increased capacity to mitigate systemic crises and promote long-term financial sustainability. However, it is important to highlight a limitation regarding the time frame in which the predictive model has been in operation. A larger dataset will be necessary to solidify the actual improvements in default rates through machine learning, specifically with the GBDT algorithm.

REFERENCES

Agarwal, M., & Vandana, T.R. (2022). Exchange rate crises in Latin America, East Asia and Russia. *Brazilian Journal of Political Economy, 42*(2), *263-282,* http://dx.doi.org/10.1590/0101-31572022-3299

Altman, E.I., Sabato, G., & Wilson, N. (2010). The value of non-financial information in small and medium-sized enterprise risk management. *The Journal of Credit Risk, 6*(2), 95-127.

Ambavat, P. P. (2021). Credit Bureaus Must Adopt AI-ML, Data Analytics for Holistic Credit Scores. *CRIF Highmark*. https://www.crifhighmark.com

Avelar, E. A., Leocádio, V. A., Campos, O. V., Ferreira, P. O. & Orefici, J. B. P. (2022). Algoritmo Random Forest para Previsão de Comportamento de Preços de Ativos. *Revista FSA, 19*(10). http://www4.unifsa.com.br/revista/index.php/fsa/article/view/2592

Banco BS2 (2024). Somos o BS2. https://www.bancobs2.com.br/somos-o-bs2/

Banco Central do Brasil (1999). *Resolução CMN n° 2.682 de 21/12/1999*. https://www.bcb.gov.br/estabilidadefinanceira/exibenormativo?tipo=Resolu%C3%A7%C3%A3o&numero=2682

Banco Central do Brasil (2022). *Resolução CMN n° 5.037 de 29/9/2022*. https://www.bcb.gov.br/estabilidadefinanceira/exibenormativo?tipo=Resolu%C3%A7%C3%A3o%20CMN&numero=5037

Banco Central do Brasil (2023). *BC e bureaus de crédito assinam acordo para compartilhamento de informações*. https://www.bcb.gov.br/detalhenoticia/668/notici

Banco Central do Brasil (BCB-CADOC) (2024). *Estatísticas monetárias e de crédito*. https://www.bcb.gov.br/estatisticas/estatisticasmonetariascredito

Bank for International Settlements (BIS) (2024, julho). *Basel Committee on Bank Supervision: International Convergence of Capital Measurement and Capital Standards*. https://www.bis.org/publ/bcbs238.htm

Berrar, D. (2019). Cross-Validation, Bootstrap, and ROC Analysis. *Encyclopedia of Bioinformatics and Computational Biology*, 542-560. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Botelho, D. ,& Tostes, F. D. (2011). Modelagem de probabilidade de churn. *Revista de Administraçäo de Empresas, 4*(396).

Calvo, G. A. (2008). Crises in Emerging Markets Economies: A Global Perspective," Central Banking, Analysis, and Economic Policies Book Series. In Kevin Cowan & Sebastián Edwards & Rodrigo O. Valdés & Norman Loayza (Series Editor) & Klaus Schmidt- (ed.), *Current Account and External Financing* (ed. 1, 12, chapter 3, 085-115), Central Bank of Chile.

Castro, C. L. de & Braga, A, P. (2011), Aprendizado supervisionado com conjunto de dados desbalanceados. *Revista Controle & Automação, 22*(5),441-466.

Creswell, J. W. & Poth, C. N. (2016), *Qualitative inquiry and research design: choosing among five approaches*. Sage.

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785

Dabrowski, M. (2023). Thirty years of economic transition in the former Soviet Union: Microeconomic and institutional dimensions. *Russian Journal of Economics, 9*,1-32. https://doi.org/10.32609/j.ruje.9.104761

Davenport, T. H., & Harris, J. G. (2007). *Competing on Analytics: The New Science of Winning*. Harvard Business Press.

Denzin, N. (1978) *The research act: a theoretical introduction to sociological methods*. (2a ed). Mc Graw-Hill.

Demirgüç-Kunt, A., & Singer, D. (2017). Financial inclusion and inclusive growth: A review of recent empirical evidence. *World bank policy research working paper*, (8040). http://documents.worldbank.org/curated/en/403611493134249446/pdf/WPS8040.pdf

Dewasiri, N. J., Dharmarathna, D. G; Choudhary,M.,(2024). Leveraging Artificial Intelligence for Enhanced Risk Management in Banking: A Systematic Literature Review. In Singh et al (Eds). *Artificial Intelligence Enabled Management: An Emerging Economy Perspective*, Chapter 13, 197-213. https://doi.org/10.1515/9783111172408013

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, *297*(3), 1178-1192.

Flick, U. (2004). *Introdução à pesquisa qualitativa*. Bookman.

Fosu, S., Boapeah, H. A. & Ciftci, N. (2023). Credit information sharing and cost of debt: Evidence from the introduction of credit bureaus in developing countries. *Financial Review, 58*(4), 653-930.

Freeman, R. E., & Dmytriyev, S. D. (2017). Corporate social responsibility and stakeholder theory: Learning from each other. *Symphonya. Emerging Issues in Management*, (1), 7-15. https://symphonya.unicusano.it/index.php/symphonya/article/view/2017.1.02freeman.dmytriyev

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics, 29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Gil, A. C. (2009). *Estudo de Caso*. Editora Atlas.

Godoy, A. S. (2006). Estudo de caso qualitativo. In C. K. Godoi, R. Bandeira-de-Mello, & A. B. Silva. *Pesquisa qualitativa em estudos organizacionais: paradigmas, estratégias e métodos.* Saraiva.

Godoi, C. K., Bandeira-De-Melo, R., & Silva, A. B. (Orgs.). (2010). *Pesquisa qualitativa em estudos organizacionais: paradigmas, estratégias e métodos* (2a ed.). Saraiva.

Grunert, J., Norden, L., & Weber, M. (2005). The role of non-financial factors in internal credit ratings. *Journal of banking & finance*, *29*(2), 509-531.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 160*(3), 523-541.

Jordan, M. I., & Mitchell, T. M. (2015). Machine Learning: Trends, Perspectives, and Prospects. *Science, 349*(6245). https://www.cs.cmu.edu/~tom/pubs/Science-ML-2015.pdf

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767-2787.

Khemakhem, S., & Boujelbene, Y. (2018). Predicting credit risk on the basis of financial and non-financial variables and data mining. *Review of accounting and finance*, *17*(3), 316-340.

Lassance, L. C. B. K., & Ternoski, S. (2021). Score ia cresol: Utilizando inteligência artificial para estimar viabilidade de crédito. *Revista Aproximação, 3*(06). https://revistas.unicentro.br/index.php/aproximacao/article/view/6923

Lee, I., & Shin, Y. J., (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons, 63*, 150-170. https://doi.org/10.1016/j.bushor.2019.10.005

Leo, M., Sharma, S., & Maddulety, K. (2019). Machine learning in banking risk management: A literature review. *Risks, 7*(1), 29.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update. *European Journal of Operational Research, 247*(1), 124-136. https://www.sciencedirect.com/science/article/pii/S0377221715007692

Louzada, F., Ara, A, Fernandes, G. B. (2016), Classification methods applied to credit scoring: Systematic review and overall comparison, *Surveys in Operations Research and Management Science, 21*, 117-134. http://dx.doi.org/10.1016/j.sorms.2016.10.001

Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition, 91,* 216-231. https://doi.org/10.1016/j.patcog.2019.02.023

Marini, J. M. & Manfrim, L. F. (2020). Metodologia de análise de crédito aplicada na redução do risco de inadimplência. *REGRAD - Revista Eletrônica de Graduação do UNIVEM, 13*(1), 76-91. https://revista.univem.edu.br/REGRAD/article/view/3105

Mashrur, A., Luo, W., Zaidi, N. A., & Robles-Kelly, A. (2020). Machine learning for financial risk management: a survey. *Ieee Access*, *8*, 203203-203223.

Mendonça, A. R. R. D., & Deos, S. (2020). Regulação bancária: uma análise de sua dinâmica por ocasião dos dez anos da crise financeira global. *Revista de Economia Contemporânea*, *24*, e202427. https://doi.org/10.1590/198055272427

Mian, A., & Sufi, A. (2009). The Consequences of Mortgage Credit Expansion: Evidence from the U.S. Mortgage Default Crisis. *The Quarterly Journal of Economics, 124*(4), 1449-1496. https://doi.org/10.1162/qjec.2009.124.4.1449

Modigliani, F., & Brumberg, R. (1954). Utility analysis and the consumption function: An interpretation of cross-section data. In K. K. Kurihara (Ed.). *Post-Keynesian Economics*. Rutgers University Press.

Montevechi, A. A., Carvalho Miranda, R., Medeiros, A. L., & Montevechi, J. A. B. (2024). Advancing credit risk modelling with Machine Learning: A comprehensive review of the state-of-the-art. *Engineering Applications of Artificial Intelligence, 137*, 109082. https://doi.org/10.1016/j.engappai.2024.109082

Moodys Local (2024). Relatório do Emissor: Banco BS2 S.A. 2024-05-13. https://moodyslocal.com.br/reporte/issuer-report/relatorio-do-emissor-banco-bs2-s-a/

Oliveira, R., & Santos, P. (2022). Avaliação de scores de crédito e práticas de bureau no Brasil. *Journal of Credit Analysis, 28*(3), 123-138. http://doi.org/10.1000/j.joca.2022.03.014

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Books.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830. https://www.researchgate.net/publication/51969319_Scikit-learn_Machine_Learning_in_Python

Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine learning*, *42*, 203-231. https://doi.org/10.1023/A:1007601015854

Rahman, M., Ming, T. H., Baigh, T. A., & Sarker, M. (2021). Adoption of artificial intelligence in banking services: an empirical analysis. *International Journal of Emerging Markets*, *18*(10), 4270-4300. https://doi.org/10.1108/IJOEM-06-2020-0724

Rahmani, A. M., Rezazadeh, B., Haghparast, M., Chang, W. C., & Ting, S. G. (2023). Applications of artificial intelligence in the economy, including applications in stock trading, market analysis, and risk management. *IEEE Access*.

Reinhart, C. M., & Rogoff, K. S. (2020). This time is different: A panoramic view of eight centuries of financial crises. *Journal of Economic Perspectives, 34*(3), 3-24. https://www.aeaweb.org/articles?id=10.1257/jep.34.3.3

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). https://doi.org/10.1145/2939672.2939778

Rizvi, S. A. R., Arshad, S., & Alam, N. (2015). Crises and contagion in Asia Pacific—Islamic v/s conventional markets. *Pacific-Basin Finance Journal, 34,* 315-326. https://doi.org/10.1016/j.pacfin.2015.04.002

Serasa Experian. (2023). *Importância do Bureau de crédito para análise de crédito*. https://www.serasaexperian.com.br/conteudos/credito/bureau-de-credito-conceito-e-importancia-na-analise-de-credito/

Sfeir, E. (2023). *Inovação aplicada ao mercado de crédito e ao setor de birôs*. https://anbc.org.br/inovacao-aplicada-ao-mercado-de-credito-e-ao-setor-de-biros/

Stake, R. E. (1995). *The art of case study research*. Sage Publications.

Stake, R. E. (2000). Case studies. In N. K. Denzin, & Y. S. Lincoln, Y. S. *Handbook of qualitative research,* (2ª ed.), 435-454, Thousand Oaks.

Sun, Y., Kamel, M. S., Wong, A. K., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern recognition*, *40*(12), 3358-3378.

Timotio, J. G. M., Vieira, V. E. L., Oliveira, R. A. de, & Silva, R. C. F. e. (2024). Inteligência Artificial no campo de finanças. *Revista de Gestão e Secretariado, 15*(6), e3935. https://doi.org/10.7769/gesec.v15i6.3935

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International data privacy law*, *7*(2), 76-99. https://doi.org/10.1093/idpl/ipx005

Wanzeller, W. F., Alves, C. M. O., & Cota, M. P. (2023). Sistema de apoio à decisão integrando cadastro negativo, scoring, análise qualitativa de crédito com inteligência artificial e criação de contratos: Protocolo para revisão de escopo. *Research, Society and Development, 12*(7), e18012742680. https://doi.org/10.33448/rsd-v12i7.42680

Zhang, T., Zhang, W., Wei, X. U., & Haijing, H. A. O. (2018). Multiple instance learning for credit risk assessment with transaction data. *Knowledge-Based Systems*, *161*, 65-77. http://dx.doi.org/10.1016/j.knosys.2018.07.030

Zhang, L., & Song, Q. (2022). Credit Evaluation of SMEs Based on GBDT-CNN-LR Hybrid Integrated Model. *Wireless Communications and Mobile Computing*, 2022. https://doi.org/10.1155/2022/5251228

Zhou, J., Li, W., Wang, J., Ding, S., & Xia, C. (2019). Default prediction in P2P lending from high-dimensional data based on machine learning. *Physica A: Statistical Mechanics and its Applications*, *534*, 122370. https://doi.org/10.1016/j.physa.2019.122370

Zöller, M. A., & Huber, M. F. (2021). Benchmark and survey of automated machine learning frameworks. *Journal of artificial intelligence research*, *70*, 409-472. https://arxiv.org/pdf/1904.12054

Yazan, B. (2016). Três abordagens do método de estudo de caso em educação: Yin, Merriam e Stake. *Meta: Avaliação, 8*(22), 149-182. http://dx.doi.org/10.22347/2175-2753v8i22.1038

Yin, R. K. (2009). *Case Study Research: Design and Methods* (4a ed.). Sage.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interest regarding this submitted work.

AUTHOR CONTRIBUTIONS

Roles	1ª author	2º author
Conceptualization	♦	♦
Data Curation	♦	
Formal Analysis		♦
Funding Acquisition		
Investigation	♦	
Methodology		♦
Project Administration	♦	♦
Resources	♦	
Software	♦	
Supervision		♦
Validation	♦	♦
Visualization	♦	♦
Writing – Original Draft	♦	
Writing – Review and Editing		♦